Reviewers' comments:

Reviewer #1: This manuscript by Wang et al. describes a novel rodent task for exploring explore-exploit decision making across varying "time horizons". Their approach is based on a similar task characterized humans in an earlier paper, and the current study includes further examination of human explore-exploit decisions on this task. The influence of relevant variables on task performance are described (though inferential statistics are scant), and this is supplemented with Bayesian model fitting to characterize parameters associated with two established heuristics for exploration: directed exploration (exploration threshold parameter) and random exploration (decision noise parameter). Like humans, rats performed the task well and were able to guide their choices by comparing recent information from guided trials with prior knowledge of possible outcomes. A major species difference was that humans increased their initial exploration of the unguided (unknown) option with time horizon (i.e., number of future opportunities to exploit feedback from guided and early choice trials), which represents a rational, directed exploration heuristic. Rats, instead, showed the reverse relationship, and additional data indicated that their exploratory behavior may have been driven in part by volatility in reward contingencies. Rats also appeared to show improved performance and less initial exploration when allowed to self-guided vs. being forced to sample one of the two response options. This is a really interesting study and the authors do a good job of highlighting the potential advantages and novel elements of this task versus other rodent tasks for measuring explore-exploit decision making. Despite some current issues (listed below), I believe the manuscript would be of interest to many in the field.  
  
1) The conceptual framework for the current study is not very well developed, particularly in the Introduction, which assumes quite a bit of readers. The main question at hand - i.e., how time horizon relates to explore/exploit decisions is not really discussed until late in the paper (in Discussion), which would make it difficult for readers to understand the purpose of experimental design parameters, as well as predictions about results. Also, much of the introduction lists limitations of existing reversal learning tasks but it is not very clear how the current approach will improve on this. For instance, without providing a clearer idea of the structure of the current task and how it relates to relevant decision making variables, it is not obvious why the current task does not have the same limitations. The current task also involves a 2-option choice with differential reward, so why is it better than reversal learning?

* We thank the reviewer for the critique on the clarity of the conceptual framework. We have expanded the description of time horizon in the introduction and elaborated on how time horizon relates to exploration. The description of time horizon is now written in a separate paragraph in the modified manuscript.
* We also modified the paragraph about the limitations of the reversal learning paradigm, and explicitly stated how our design addressed these limitations in the last paragraph of the introduction. Our task is better than the reversal learning approach mainly in the following ways:
  + Firstly, both good and bad outcomes should occur in exploration. However, in reversal learning, after the reversal point, “exploring” the previously suboptimal option will always lead to a better outcome. In reversal learning paradigms, exploration is confounded with simply abandoning a bad option, while in our design, exploring the unguided option can lead to either better or worse outcomes.
  + Secondly, it is notoriously difficult to separate different drives and heuristics for exploration in reversal learning paradigms. To study directed exploration (i.e. uncertainty driven exploration) for example, there needs to be a difference in uncertainty between the two options. For reversal learning, this uncertainty difference is implicit in that the less chosen option has more uncertainty. Since the less chosen option in reversal learning usually also has a lower estimated value, value and uncertainty are confounded in reversal learning. However, in our design both value and uncertainty are manipulated independently from each other, allowing us to dissociate uncertainty from value and properly measure directed exploration.

2) There are very few statistical results provided. Instead, the Results consists of general descriptions of group level performance without evidence of the significance of findings. This is generally problematic but especially for subtle effects (e.g., constant vs. random reward differences in Fig 10; horizon effects on model parameter estimates in Figure 8; reference to horizon effect on switch in Fig 4D on p. 14).

We conducted additional statistical analyses and included more details. We added missing p-values throughout the results section of the manuscript. We re-ran all the analyses after incorporating suggestions from other comments of the reviewer (e.g. controlling for feeder preference). Here we highlight the main findings of the paper with statement of statistical significance.

* + - 1. (Similar to humans) Rats were able to use prior information to guide exploratory choices.

In figure 4C, rats were able to choose the high reward option at the first free choice (without knowing the reward of the unguided option) significantly above chance (p < 0.001 for H = 1 and 6, p = 0.01 for H = 15). Rats can only perform above chance if they have access to prior information.

* + - 1. (Similar to humans) Rats adapted the extent to which they explore based on the guided reward size. In figure 7C, Two-way ANOVA (Horizon x Guided reward) revealed a significant main effect of guided reward (p < 0.001).
      2. (Different than humans) Rats used less directed exploration (i.e. have lower thresholds) in long horizons.
         1. In Experiment 1(Figure 8), there is a significant main effect of horizon on the threshold parameter for humans (p < 0.001), but there is no significant effect of horizon on the threshold parameter for rats (p > 0.05). In other words, when different horizon conditions are tested between sessions, we didn’t find a significant horizon effect on directed exploration in rats.
         2. In Experiment 2(Figure 9), there is a significant main effect of horizon on the threshold parameter for rats (p < 0.001). There is also a significant main effect of horizon on the model-free p(unguided) measure (p = 0.003). Together, these results showed a significant horizon effect on directed exploration if rats experienced both horizons within the same session. (Note that the horizon conditions are always within-session for human experiments.)
      3. (Rats alone) Rats explored differently in self-guided exploration compared to cue-guided exploration. In Figure 11D, we showed that there was a significant main effect of nGuided (0 for self-guided vs 1 for cue-guided) on p(unguided), p < 0.001.

Addressing the particular figures that the Reviewer mentioned:

* + - 1. (Figure 10D, F) For random reward condition, we showed that there was a significant increase of both threshold (p < 0.01) and noise (p < 0.01) parameters compared to the constant reward condition.
      2. (Figure 8) See 3a above. We have added to figure 8 that the effect was non-significant.
      3. (Figure 4D) P(switch) at later trials (trial# 2-6 in H = 6 vs trial# 11-15 in H = 15) is significantly lower for H = 15 than for H = 6 (p < 0.001).

3) Some additional methodological details should be provided or clarified. What volume were the sugar water drops? Were the rats food/water restricted? How many trials/games per session? Were rats trained through different phases at different rates based on performance? Were rewards ever symmetrical across options within games and if so how was this dealt with for analysis?

* The volume of sugar water drop was 150 microliter per drop.
* The rats were food restricted to 85% of their ad libitum weight and were not water restricted.
* For Experiment 1, rats on average do 31.3 games and 125.3 trials per session for H = 1, they do 13.5 games and 121.5 trials per session for H = 6, and an average of 6.5 games and 123.4 trials for H = 15. On average, each rat completed 348.7 H = 1 games, 375.8 H = 6 games, and 170 H = 15 games.

For Experiment 2, rats completed an average of 25.62 games and 130.44 trials per session.

* Rats were pretrained through different phases at different rates during pretraining. We first trained rats to associate light with reward, then they were trained to go to the homebase to trigger the lights at the reward feeders (first with reward at homebase, then without the homebase reward), then they were trained to learn that two feeders give different amounts of rewards (first 0 vs 1 drops, then 1 vs 5 drops, then the full reward schedule).Rats went through these phases of pretraining at different rates based on their individual performance. After pretraining, all rats performed three experiments in the order: Experiment 1, Experiment 2 and then Experiment 3. We have added a paragraph describing this pre-training protocol in the manuscript.
* Rewards of the two options were generated independently using a custom written MATLAB program for each game. So indeed, rewards for two options can be identical (6.6% of all trials). These trials were included in all analysis that focused on the 1st free choice (since rats only know the guided reward before making the first free choice, whether the unguided option has an identical value did not matter). These trials were excluded when doing analysis of later trials, as in Figure 4, 5, 10AB, 11AB.

4) For Experiment 2, it was unclear exactly how H conditions were organized within session, though it seems to be the case that they were strictly alternated when confounded with home base. This should be in Methods. Also for this experiment, how were nG 0, 1, and 3 conditions organized (e.g., blocks of sessions, randomly across sessions).

* Horizons were strictly alternated in Experiment 2, for a given session, one of the homebase was always used for H = 1 and the other always for H = 6. For each session, a MATLAB program pseudo-randomly made the homebase/horizon condition pairing. Consequently, Homebase A could be H = 1 for one session and H = 6 for another session.
* nG = 0,1,3 were run in blocks of sessions.

5) For the human task, does the schematic in Figure 3 represent actual task stimuli and procedures? For example, were subjects given a tally of past reward histories for all trials within a game? This should be indicated in Methods.

* The schematic in Figure 3 represents the actual task stimuli. Subjects were indeed given a tally of past reward histories for all trials. We have added this information to the Methods.

6) Do the data presented in figures represent all the data from all relevant sessions or were they restricted to sessions after rats had time to learn about the new task contingencies? For example, rats presumably took some time to learn about the change in time horizon across blocks of sessions in Experiment 1. And the same goes for when they switched to the within-session analysis of horizon in Experiment 2, and the random task in Experiment 3. As noted by the authors, performance in the random task shows some clear carryover from the earlier phases of testing. This should be specified as good practice but also raises questions about species differences.

* We thank the reviewer for this useful suggestion. The data presented in the initial submission used all the data from all relevant sessions. In the current submission, we adopted the reviewer’s suggestion and analyzed the data after excluding the transitioning sessions. In Experiment 1, we excluded 1 whole session after each transition of horizon conditions. In Experiment 2, we excluded the first session for each rat, and we excluded the first 2 games in each session (Although we have high/low pitch sound cues played at the homebases to help signal H = 6/H = 1 horizon condition respectively, in practice, it may take rats 1 full game to learn the associated horizon condition with each homebase). These exclusions did not change our results or conclusions.
* Experiment 3 was carried out at the end of Experiment 2 to test how volatility might account for horizon-dependent changes in exploration in rats. Despite the possible carry-over effects (we expect carry-over effects for threshold but not for decision noise), the fact that volatility changes both the decision noise parameter and the threshold parameter, and that only threshold changes across horizons in actual behavior, suggest that volatility may not be what is driving the threshold specific horizon-dependent changes in rats.

7) Humans seem to have little trouble deploying a directed exploration strategy based on time horizon and guided choice feedback (reward size). They seem to explore the unguided option during the first free choice if there is any question about what the best option is, particularly when there is a long horizon to exploit that information. The authors state (p. 10) while showing similar early exploration, "it took longer for rats to switch back," referring to their persistent switching behavior in Fig 4 and 5. But these data are really able to get at this question precisely because they don't describe whether the switch is moving away or toward the guided choice. Given the rats' generally poorer performance and persistent tendency to switch within games, even with long horizons, suggests that they were switching back and forth from the best option. This is later discussed in the context of the random reward task, but the authors should avoid giving the wrong impression when discussion Fig 4 and 5.

This is a very good point. To address the reviewer’s comment, we split p(switch) into p(switch away) vs p(switch back). We found that the difference between the “guided = good” option and the “guided = back” option occurred in the direction of switching back to the guided choice. On top of the persistent baseline switching in rats, rats do switch back from the unguided option to the guided option more (when the guided option is objectively better) up to trial #4 in H= 6, p = 0.01. A new figure was added.

  
8) The cross species comparisons are a bit strained despite the general similarities across tasks. For instance, humans appear to receive a continuous tally of past reward within each game, which explains their lack of later exploration once they understand the basic task. But rats could forget this information and decide to re-explore the options, particularly in games with long horizons though reward volatility on short horizon games may provide a separate reason to explore.

We thank the reviewer for bringing up this critical point. We want to point out the following:

We agree that despite the matching underlying structure, there are some differences between the rat and human version of the task (reward history vs no history, points for humans and juice for rats, effortful spatial runs in rats vs effortless key presses in humans). However, our main interest is in how humans and rats change their exploration behavior across horizon conditions. This horizon comparison is done within species (We know of no evidence that differences in the physical implementation of the task would contribute to the horizon difference within species), and then we showed a qualitative difference between species that humans increase whereas rats decrease the threshold parameter in long horizon condition compared to short horizon condition. Given that the horizon difference is calculated in a similar manner within species, the results should be valid within each species. This being said, our results are indeed prompting further work and new tasks in humans that would match that of the rats, for example a physical spatial task where human subjects walk from home base to the bandits. These new tasks are left for future work.

The reviewer is correct that rats might re-explore because of forgetting. For the same number of guided trials, rats are equally likely to forget in H = 1 and H = 6 however. We agree that reexploring is an important form of exploration (uncertainty increases as animals forget, and reexploring can be considered a form of directed exploration), but forgetting does not predict a difference in behavior across horizon conditions.

9) The authors suggest that rats may be performing the task to satisfice instead of optimize, like their human counterpart. This is related (p. 17) to the higher tendency for humans to explore than rats. This conclusion is partly based on the idea that rats must exert more effort and may be less willing to explore (a point also made earlier in the manuscript). But of course this specific data refers to first trial exploration. Rats were not less willing to explore on long horizon games. As indicated above, they tended to switch back and forth throughout the game (i.e., even when they had sampled both reward contingencies).

This is a really good point. We agree that the difference in effort can not account for the fact that rats were willing to switch in later trials in long horizon games. This constant switching in later trials may indeed reflect random exploration. We agree it was not completely accurate to say that “more effort” leads to “being less willing to explore”. In the modified manuscript, we emphasize that “more effort” relates more to “not optimize” rather than “not explore”. It changes the relative utility/effort balance of the options, so that it is not worth risking getting a 1 drop (to optimize the check if the unguided reward is a 5) if the guided reward is 4 drops for rats. This deliberate way of exploration is directed exploration rather than random exploration. “Effort” might influence directed exploration more than random exploration. To properly test the effect of efforts on exploration, future experiments could potentially run identical tasks in long/short distance (or maze vs boxes). We would predict that rats in an effortful setup will be less willing to engage in directed exploration than rats operating in an effortless setup (e.g. a lever pressing task). We re-wrote the relevant sections.

10) The greater preference for high reward option and lower level of switching in the free choice vs. guided choice is interesting but one account not considered is that rats some feeder/spatial preferences that bias them on these free choice first trials that continuous to bias their performance to the same degree within each game. This make sense generally and also explains why the effect of free vs. guided choice is so stable across variables like guided reward size and horizon. This could be a long-term feeder bias or perhaps something more dynamic (e.g., like a preference for whichever feeder has been paying off better in recent sessions).

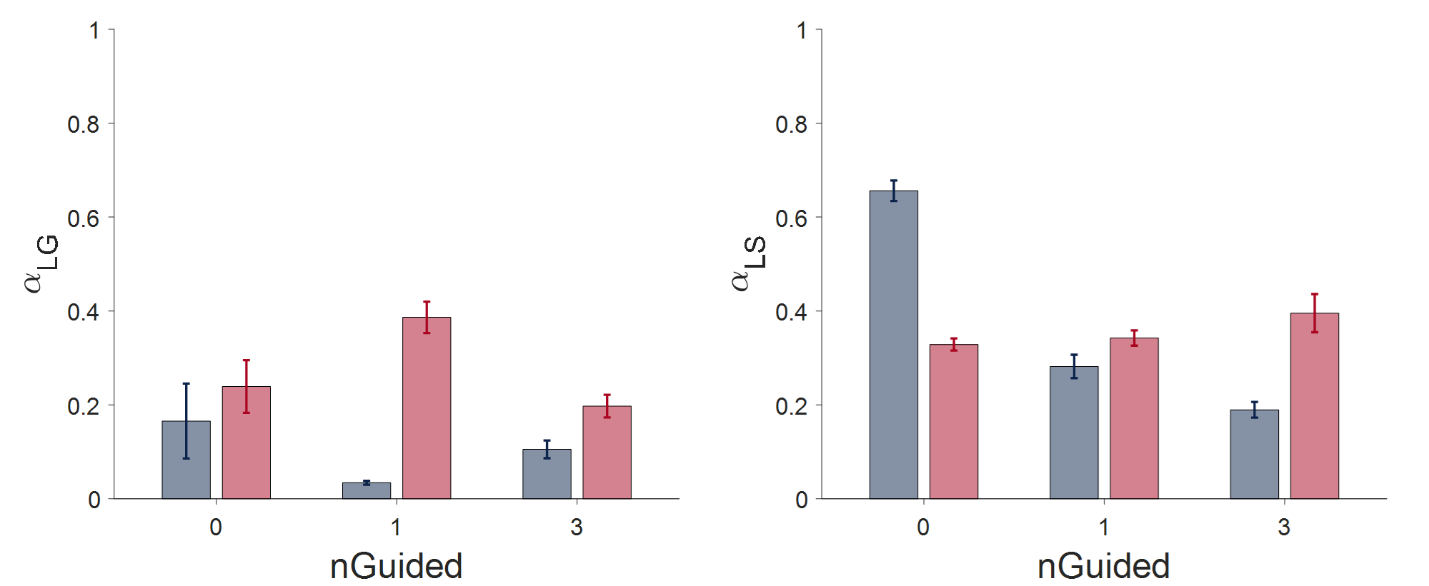
We thank the reviewer for raising this critical and insightful point. In an attempt to measure the influence of short-term/long-term feeder bias, we calculated two measures. We first computed the reward from last game(most recent reward of the current feeder from a previous game within session) to quantify short-term feeder bias, and computed the average reward from last session to quantify long-term feeder bias.



We first plot p(explore) as a function of both (NaN means that the rat didn’t choose the guided feeder in the previous game) and (LS values are binned). ANOVA analysis showed that both LG (p < 0.001) and LS (p = 0.02) have significant influence on p(explore). These showed that rats do have feeder biases. (Humans do not.)

As a result, we repeated all the model-based analysis by including these two additional parameters in the model to account for feeder biases. Specifically, we now have

Here, in addition to the current reward, decision threshold, and spatial bias, we added two new terms, the first LG term quantifying the influence of feeder bias from last game, and the second LS term quantifying the influence of feeder bias from last session. Parameter recovery shows that our model estimates the upper bound of the feeder bias coefficients (Our model tends to overestimate , Fig S1). Through model fitting, we confirm and acknowledge that rats are influenced by both short-term and long-term feeder bias. LG coefficient is significantly larger in H = 6 than H = 1 condition, showing that short term feeder bias (from last game) has a significantly bigger influence on H = 6 games (p < 0.001). This is likely due to the fact that rats spend more trials at H = 6 feeders within a session. There are no differences in long term feeder bias (from last session) between horizon conditions (p = 0.48).



(We note that the coefficients are relative. appears to be larger than only because , in fact, has roughly the same impact on choices as .)

Next, we confirmed that after accounting for feeder biases, the model-based results in the paper still holds, e.g., we still see reliable differences in threshold between horizon conditions (All the figures have been re-built using this extended model) in Experiment 2. In particular, we still observe the difference in the threshold parameter between self-guided vs cue-guided trials.



Despite these feeder biases, our model suggests that the large difference in thresholds between self-guided vs cue-guided conditions is still present. These figures and results were added in supplemental materials.

11) On a related note, a spatial bias parameter was computed in the model fitting but was not discussed.

The results of spatial bias parameter are now added to the supplementary materials. The spatial bias is centered at 0 in all conditions in Experiment 1.



But we did observe a left side bias when rats were guided only once (nG = 0 and 1) in Experiment 2. Interestingly, this left side bias is compatible with previous work from our laboratory in a different spatial task relevant to spatial navigation optimization, and may be related to rat right-handedness (Watkins de Jong, 2011).

The Traveling Salesrat: Insights into the Dynamics of Efficient Spatial Navigation in the Rodent. Watkins L., Gereke B. G. M. Martin, JM Fellous . J Neural Engineering, 8(6), 2011.

Chart, histogram

Description automatically generated

The bias does not change significantly with horizon (p > 0.05).

Minor issues:  
12) P. 10 - statement about boredom or motor error seems to refer to residual responding at last trial but worded as if about the reason for the decrease in switching.

We thank the reviewer for pointing this out. We have rephrased this on P.10.

13) P. 11 - it says "3 or 4 drops for humans"

We have rephrased this on P.11